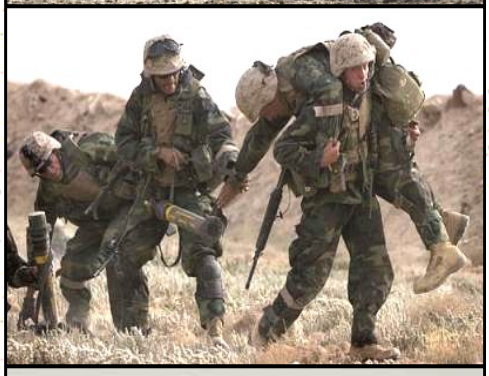

A Comparison of The PRIME-MDPHQ-9 and PHQ-8 in A Large Military Prospective Study, The Millennium Cohort Study

Timothy S. Wells
Jaime L. Horton
Cynthia A. LeardMann
Isabel G. Jacobson
Edward J. Boyko

Naval Health Research Center

Report No. 12-20

This research was conducted in compliance with all applicable federal regulations governing the protection of human subjects in research.

Naval Health Research Center
140 Sylvester Road
San Diego, California 92106-3521

Journal of Affective Disorders 148 (2013) 77–83

Contents lists available at SciVerse ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research report

A comparison of the PRIME-MD PHQ-9 and PHQ-8 in a large military prospective study, the Millennium Cohort Study

Timothy S. Wells [a], Jaime L. Horton [a,*], Cynthia A. LeardMann [a], Isabel G. Jacobson [a], Edward J. Boyko [b]

[a] Deployment Health Research Department, Naval Health Research Center, San Diego, CA, USA
[b] Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Puget Sound Health Care System, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 24 July 2012
Accepted 20 November 2012
Available online 14 December 2012

Keywords:
Depression
Epidemiology
Assessment/diagnosis
Measurement/psychometrics
Mood disorders

ABSTRACT

Background: In light of increased concerns about suicide in the military, institutional review boards have mandated increased scrutiny of the final item on the depression screening tool, the PHQ-9, which asks about suicidal thoughts. Since real-time monitoring of all individual responses in most observational studies is not feasible, many investigators have adopted the PHQ-8, choosing to remove the ninth item. This study compares the performance of the PHQ-8 with the PHQ-9 in a population-based sample of military or nonmilitary subjects.
Methods: The Millennium Cohort Study administers a self-reported questionnaire that includes the PHQ-9 at 3-year intervals to current and former U.S. military personnel. PHQ-9 responses of 143,705 Millennium Cohort members were investigated. Cross-sectional comparisons of the PHQ-9 and PHQ-8 and prospective analyses to detect a 5-unit change in these measures were performed.
Results: Greater than substantial agreement was found between the PHQ-8 and 9 instruments (kappas, 0.966–0.974 depending on survey cycle). There was similarly high agreement between the PHQ-8 and 9 in detecting a 5-point increase (κ=0.987) or decrease (κ=0.984) in score.
Limitations: One potential limitation of this study is that participants completed the PHQ-9, and PHQ-8 scores were extrapolated from the PHQ-9. In addition, the Millennium Cohort may not fully represent the U.S. military; though previous evaluations have shown the cohort to be a well-representative sample.
Conclusions: Since excellent agreement was detected between the PHQ-8 and PHQ-9 instruments, the PHQ-8 would capture nearly all the same cases of depression as the PHQ-9 in populations similar to the one in this study.

1. Introduction

The Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PHQ) is a standardized instrument that provides an assessment of mental health status based on scores of several health concepts (Spitzer et al., 1999, 2000, 1994). A 9-item scale from this instrument (PHQ-9) used to screen for major depressive disorder (Spitzer et al., 1999), has been shown to have high sensitivity (0.93) and specificity (0.89) (Fann et al., 2005), and correlates well with a diagnosis of depression, as outlined in the Diagnostic and Statistical Manual of Mental Disorders (2000). The ninth and final item on the scale asks about thoughts of being better off dead or hurting oneself, which is known to be related to suicidal thoughts and ideation (Corson et al., 2004). Exclusion of the final item results in the PHQ-8, which has also been shown to be a valid instrument for evaluating depression symptoms in specific populations (Kroenke et al., 2009).

The Millennium Cohort Study (Ryan et al., 2007; Smith, 2009), the largest population-based longitudinal cohort study in military history, has included the PHQ-9 as part of the standard questionnaire since the study launched in 2001. Symptoms of depression have previously been examined in this population using the PHQ-9 (Ryan et al., 2007; Wells et al., 2010), with positive screens for new-onset depression occurring in approximately 4% of men, and 8% of women at the time of the first follow-up questionnaire (Wells et al., 2010). However, in light of increased concerns about suicide in the military (Kuehn, 2009; Oquendo et al., 2005), institutional review boards have mandated increased scrutiny of this item with the intent of initiating provider referral when respondents positively endorse it. In nonclinical research settings, many investigators have adopted the PHQ-8, choosing to remove the final item that may indicate suicidal thoughts. Since real-time monitoring of more than 150,000 responses from Millennium Cohort members is not feasible, starting in 2011 the final item

* Corresponding author. Tel.: +1 619 767 4905; fax: +1 619 553 7601.
E-mail address: jaime.horton@med.navy.mil (J.L. Horton).

0165-0327/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jad.2012.11.052

was removed from the questionnaire and the study currently uses the PHQ-8 to screen for depression. Although the PHQ-8 has been shown to have similar operating characteristics as the PHQ-9 (Kroenke and Spitzer, 2002), this was done using a sample of 6000 individuals seeking treatment in primary care or obstetrics-gynecology clinics (Kroenke et al., 2001). To our knowledge, no study has validated the PHQ-8 in a large, population-based cohort, military or otherwise. The objective of this study was to compare the PHQ-8 with the PHQ-9 to understand differences in depression screening capability between these two instruments in the Millennium Cohort Study, which may be generalizable to other similar population-based studies. Unique features of this study include investigating differences in a large military population that is probably healthier than a general population sample. Also, with the large sample size, subgroups such as differences by sex can be explored to determine performance of the PHQ-8 compared with the PHQ-9.

2. Methods

2.1. Study population

The Millennium Cohort Study began collection of self-reported health outcome and exposure data in 2001, prior to the start of the operations in Iraq and Afghanistan. The Millennium Cohort currently includes over 150,000 U.S. service members who enrolled during three separate cycles (panels) between 2001 and 2008. With the goal of evaluating long-term health outcomes related to military service, participants are surveyed every 3 years throughout a 21-year planned follow-up period. Detailed descriptions of methodology for the Millennium Cohort Study have been published elsewhere (Ryan et al., 2007; Smith, 2009; Gray et al., 2002).

The first invited panel of the Millennium Cohort Study consisted of a weighted random sample of U.S. military personnel serving in October 2000, with oversampling of women, service members previously deployed to Bosnia, Kosovo, or Southwest Asia, and Reserve and National Guard members. Of the 77,047 participants who consented and enrolled, 55,021 (71%) completed the first follow-up questionnaire from 2004–2006, and 54,790 (71%) completed the second follow-up from 2007–2008. A second panel was randomly selected from military personnel with 1–2 years of service as of October 2003, with oversampling again performed for women and Marine Corps members. The second panel enrolled 31,110 consenting members from 2004–2006, 17,152 (55%) of whom completed a follow-up questionnaire from 2007–2008. Also, a third panel of 43,440 participants with 1–3 years of service as of October 2006 and oversampling of women and Marine Corps members was enrolled in 2007–2008. Demographic and military-specific data obtained from electronic personnel files include sex, birth date, highest education achieved, marital status, race/ethnicity, deployment experience in support of the operations in Iraq and Afghanistan, pay grade, service component, service branch, and duty occupations. Data for education level, marital status, deployment experience, and occupations were supplemented by self-reported data when missing from electronic files. Participants were excluded if they were missing information from the PHQ or other covariate information.

2.2. Depression

Depression was investigated using the PHQ (Spitzer et al., 1999), which is embedded in the Millennium Cohort questionnaire and provides a psychosocial assessment based on scores of several health concepts (Spitzer et al., 2000, 1994; Kroenke and Spitzer, 2002). Using a 4-point Likert scale, participants rated the severity of each depressive symptom from "not at all" to "nearly every day" during the 2 weeks prior to questionnaire completion, where a higher score indicates greater severity (Spitzer et al., 1999). Using the PHQ 9-item scale, participants screened positive for depression if they met the following two criteria: (1) responded "more than half the days" or "nearly every day" to at least five of the nine depressive items, with "thoughts that you would be better off dead or of hurting yourself in some way" being counted if present at all, and (2) one of the items endorsed is having depressed mood or anhedonia (Kroenke and Spitzer, 2002). Similarly, using the PHQ-8, which does not include the item "thoughts that you would be better off dead or of hurting yourself in some way," individuals screen positive for depression if five or more of the eight depressive symptom criteria have been present "more than half the days" or "nearly every day" in the past 2 weeks, and one of the symptoms is depressed mood or anhedonia (Kroenke et al., 2009). Using both the PHQ-8 and the PHQ-9 instruments, participants were categorized as screening positive or negative for depression. Those who screened positive on the PHQ-9 and PHQ-8 and those who screened negative on the PHQ-9 and PHQ-8 were termed "concordant positive" and "concordant negative," respectively, while the discordant group was termed the PHQ-9 positive/PHQ-8 negative group for the purposes of this study.

2.3. Covariates

Variables considered for this study mirrored those considered for the models in the first depression study conducted using Millennium Cohort data (Wells et al., 2010). Covariates included sex, birth year, education, marital status, race/ethnicity, military pay grade, branch of service, service component, occupational category, smoking status, alcohol-related problems (Yes if one or more items endorsed on the CAGE questionnaire) (Dhalla and Kopec, 2007), a positive screen for posttraumatic stress disorder (PTSD) assessed using the 17-item PTSD Checklist-Civilian Version (PCL-C) (Weathers et al., 1993), deployment experience with or without combat, and cumulative days deployed. All independent variables were evaluated at baseline. Defense Manpower Data Center (DMDC) provided deployment data to classify individuals as nondeployed and deployed with or without combat experience. Participants were considered deployed in support of the operations in Iraq and Afghanistan if they completed at least one deployment prior to their baseline questionnaire. Combat experience was determined by an affirmative response to ever witnessing at least one of the following items on the baseline questionnaire: a person's death due to war, instances of physical abuse, dead and/or decomposing bodies, maimed soldiers or civilians, or prisoners of war or refugees. Cumulative days deployed were assessed by calculating the number of days each participant was deployed prior to baseline using in and out of theater dates from data provided by DMDC.

2.4. Statistical analyses

Univariate analyses were used to describe frequencies and proportions of those who screened positive on both instruments, those who screened negative on both instruments, and those who screened PHQ-9 positive but PHQ-8 negative. To determine the degree of nonrandom agreement between the PHQ-9 and PHQ-8, the kappa statistic (Cohen, 1960) was calculated between the two instruments for each survey period (2001–2003, 2004–2006, and 2007–2008), providing a cross-sectional comparison of performance of the PHQ-9 versus the PHQ-8 at each survey cycle.

A kappa ranging in value between 0.8 and 1.0 was defined as "greater than substantial agreement," between 0.6 and 0.8 as "substantial agreement," between 0.4 and 0.6 as "moderate agreement," between 0.2 and 0.4 as "fair agreement," and between 0.0 and 0.2 as "slight or poor agreement" (Landis and Koch, 1977). Additionally, the sensitivity of the PHQ-8 was calculated using the PHQ-9 as the "gold standard". Note that it was not possible to have a positive PHQ-8 and a negative PHQ-9, so specificity remained 100% since there were no false positives.

Kappa statistics were also used to calculate agreement between the ability of the PHQ-8 and PHQ-9 to detect a 5-point or greater change in score from baseline to first follow-up. This analysis was conducted because a 5-point decrease in the PHQ-9 has been used as a general indicator of clinically significant improvement (Kroenke and Spitzer, 2002; Kroenke et al., 2010) and it would be problematic if the PHQ-8 were not able to detect such a change. Specifically, the difference between the total score at baseline and follow-up was calculated using the PHQ-9 and then again using the PHQ-8. Each item is scored with 0–3 points; therefore, not endorsing an item at baseline and then endorsing it at follow-up can add up to 3 points, or decrease the score by up to 3 points if the converse occurs. Thus, we investigated the occurrence of the PHQ-8 not detecting a 5-point or greater change in score when the PHQ-9 detected a change and the reverse. This analysis was restricted to participants who completed at least one follow-up questionnaire, with the first available follow-up used for panel 1 participants. Panel 3 members were not included because they had yet to complete a follow-up questionnaire. The included participants were given dichotomous indicators (Yes, No) for a 5-point or greater decrease or increase using the PHQ-8 and PHQ-9 scores between baseline and follow-up. To determine if the PHQ-8 had similar characteristics to detect a 5-point increase or decrease, we again calculated kappa statistics, as well as sensitivity and specificity.

Finally, multivariable logistic regressions using cross-sectional baseline data were performed to examine associations of screening PHQ-9 positive but PHQ-8 negative compared separately with participants who were concordant positive and concordant negative. Wald chi-square statistical tests were used to test associations at the $\alpha=0.05$ level of significance between a discordant depression screen and each covariate described. Regression diagnostics were conducted to identify multicollinearity using a variance inflation factor (VIF) of 4 or greater to indicate a potential problem (Glantz and Slinker, 1990). The two deployment-related variables were found to be collinear (VIF=4.51); thus, we removed "cumulative days deployed" from all models. All other covariates were included in the multivariable models. Data management and statistical analyses were performed using SAS software, version 9.2 (SAS Institute, Inc., Cary, North Carolina).

3. Results

Of the 151,569 participants who completed at least one Millennium Cohort questionnaire, 7864 were excluded from this study due to missing data. Thus, the study population consisted of 143,705 participants (94.8%). At baseline, 6531 (4.5%) participants screened positive for depression using the PHQ-9. Of those, nearly all screened positive for depression using the PHQ-8 ($n=6163$, 94.4%) (Table 1).

Participants in the concordant positive and the PHQ-9 positive/PHQ-8 negative groups had more similar characteristics than the concordant negative group (Table 1). Notably, a higher proportion of participants who screened positive for depression based on the PHQ-9 and screened positive or negative using the PHQ-8 were younger, less educated, not married, enlisted, Army or Marine Corps members, serving on active duty, and current smokers compared with those who screened concordant negative for depression. A higher proportion of those who screened positive by the PHQ-8 or PHQ-9 also screened positive for

Table 1
Baseline characteristics of Millennium Cohort participants by PHQ-9 and PHQ-8 status ($N=143,705$).

Baseline characteristics	Concordant negative PHQ-9 and PHQ-8 $n=137,174$		Concordant positive PHQ-9 and PHQ-8 $n=6163$		PHQ-9 positive and PHQ-8 negative $n=368$	
	n	(%)	*n*	(%)	*n*	(%)
Sex						
Male	95,055	(69.3)	3,698	(60.0)	252	(68.5)
Female	42,119	(30.7)	2,465	(40.0)	116	(31.5)
Birth cohort						
Pre-1970	45,176	(32.9)	1,135	(18.4)	63	(17.1)
1970–1979	39,359	(28.7)	1,782	(28.9)	105	(28.5)
1980 and later	52,639	(38.4)	3,246	(52.7)	200	(54.3)
Education						
Some college or less	107,901	(78.7)	5,668	(92.0)	333	(90.5)
Bachelor's or higher	29,273	(21.3)	495	(8.0)	35	(9.5)
Marital status						
Currently married	67,121	(48.9)	2,465	(40.0)	126	(34.2)
Not currently married	70,053	(51.1)	3,698	(60.0)	242	(65.8)
Race/ethnicity						
White non-Hispanic	97,839	(71.3)	4,219	(68.5)	259	(70.4)
Black non-Hispanic	16,714	(12.2)	863	(14.0)	49	(13.3)
Other	22,621	(16.5)	1,081	(17.5)	60	(16.3)
Military pay grade						
Enlisted	112,246	(81.8)	5,844	(94.8)	338	(91.8)
Officer	24,928	(18.2)	319	(5.2)	30	(8.2)
Service branch						
Army	59,973	(43.7)	3,490	(56.6)	200	(54.3)
Navy/coast guard	24,885	(18.1)	1,083	(17.6)	77	(20.9)
Air force	40,388	(29.4)	849	(13.8)	41	(11.1)
Marines	11,928	(8.7)	741	(12.0)	50	(13.6)
Component						
Active duty	87,338	(63.7)	4,429	(71.9)	244	(66.3)
Guard/reserve	49,836	(36.3)	1,734	(28.1)	124	(33.7)
Occupation						
Combat specialist	24,624	(18.0)	1,058	(17.2)	78	(21.2)
Health care	14,731	(10.7)	605	(9.8)	41	(11.1)
Other	97,819	(71.3)	4,500	(73.0)	249	(67.7)
Smoking status						
Never smoker	78,996	(57.6)	2,497	(40.5)	174	(47.3)
Past smoker	31,454	(22.9)	1,510	(24.5)	70	(19.0)
Current smoker	26,724	(19.5)	2,156	(35.0)	124	(33.7)
CAGE/alcohol[a]						
No	111,464	(81.3)	3,975	(64.5)	214	(58.2)
Yes	25,710	(18.7)	2,188	(35.5)	154	(41.8)
Baseline PTSD[b]						
No	132,522	(96.6)	2,242	(36.4)	181	(49.2)
Yes	4,652	(3.4)	3,921	(63.6)	187	(50.8)
Deployment experience[c]						
Nondeployed	108,767	(79.3)	4,326	(70.2)	261	(70.9)
Deployed without combat	11,513	(8.4)	327	(5.3)	27	(7.3)
Deployed with combat	16,894	(12.3)	1,510	(24.5)	80	(21.7)
Total days deployed						
0	108,767	(79.3)	4,326	(70.2)	261	(70.9)
1–180	11,274	(8.2)	490	(8.0)	25	(6.8)
181–270	6,862	(5.0)	444	(7.2)	30	(8.2)
271 or more	10,271	(7.5)	903	(14.7)	52	(14.1)

DSM-IV-TR, Diagnostic and Statistical Manual of Mental Disorders, fourth Edition, Text Revision; PCL-C, PTSD Checklist-Civilian Version; PTSD, posttraumatic stress disorder.

[a] At baseline, participant self-reported ever feeling at least one of the following: (1) need to cut back on drinking, (2) annoyed at anyone who suggested to cut back on drinking, (3) guilty about drinking, and (4) a need for an "eye-opener," or early morning drink.

[b] Participants who screened positive for PTSD symptoms based on PCL-C and DSM-IV-TR criteria at baseline.

[c] Participants who had been deployed and reported witnessing death, trauma, injuries, prisoners of war, or refugees at baseline, were considered to have combat-associated experiences.

2001–2003 Survey Period

		PHQ-9 Positive		
		Yes	No	κ = 0.966
PHQ-8	Yes	2,233	0	Sensitivity = 93.7%
Positive	No	150	71,171	Specificity = 100%

2004–2006 Survey Period

		PHQ-9 Positive		
		Yes	No	κ = 0.972
PHQ-8	Yes	3,197	0	Sensitivity = 94.8%
Positive	No	174	77,264	Specificity = 100.0%

2007–2008 Survey Period

		PHQ-9 Positive		
		Yes	No	κ = 0.974
PHQ-8	Yes	5,208	0	Sensitivity = 95.2%
Positive	No	264	104,295	Specificity = 100.0%

Ability to Detect 5-Point Decrease

		PHQ-9 5-Point Decrease		
PHQ-8		Yes	No	κ = 0.984
5-point	Yes	5,157	17	Sensitivity = 97.4%
Decrease	No	135	65,576	Specificity = 100.0%

Ability to Detect 5-Point Increase

		PHQ-9 5-Point Increase		
PHQ-8		Yes	No	κ = 0.987
5-Point	Yes	6,320	23	Sensitivity = 97.9%
Increase	No	133	64,409	Specificity = 100.0%

Fig. 1. Cross-tabulation comparison of the PHQ-8 with the PHQ-9 by survey period and ability of the PHQ-8 and the PHQ-9 to detect a 5-point increase or decrease in score.

CAGE/alcohol, PTSD, deployed with combat experience and had deployed more total days compared with those who screened concordant negative for depression.

Three kappa analyses were performed to assess the agreement for screening positive for depression using the PHQ-8 compared with the PHQ-9 using data from three different survey waves (2001–2003, 2004–2006, and 2007–2008) and are displayed in Fig. 1. There was greater than substantial agreement ($\kappa = 0.97$) for all survey waves with high sensitivity (94–95%). A separate kappa analysis that assessed the agreement to detect a 5-point or more increase or decrease between survey waves observed very high agreement ($\kappa = 0.98$–0.99, sensitivity 97–98%, and specificity 100%) when using the PHQ-8 compared with the PHQ-9.

Table 2 shows the results from the multivariable logistic regressions representing the odds of screening PHQ-9 positive/PHQ-8 negative compared with the two concordant groups. Compared with those who screened concordant negative, the discordant group were significantly more likely to be born in 1980 or later (adjusted odds ratio [OR] = 1.63, 95% confidence interval [CI], 1.15–2.31), screen positive for CAGE/alcohol (OR = 1.74, 95% CI, 1.39–2.17), and screen positive for baseline PTSD (OR = 21.55, 95% CI, 17.24–26.93), while significantly less likely to be serving in the Air Force (OR = 0.51, 95% CI, 0.36–0.73). Compared with those who screened concordant positive, participants who screened PHQ-9 positive/PHQ-8 negative were significantly more likely to be Reserve/National Guard members (OR = 1.47, 95% CI, 1.15–1.89) and screen positive for CAGE/alcohol (OR = 1.39, 95% CI, 1.11–1.74), while they were less likely to be female (OR = 0.65, 95% CI, 0.51–0.83), a past smoker (OR = 0.65, 95% CI, 0.48–0.87), and to screen positive for baseline PTSD (OR = 0.58, 95% CI, 0.47–0.72).

4. Discussion

Clearly, the PHQ-9 is the preferred instrument of use in a clinical setting. The objective of this research was to determine how screening results may differ for the PHQ-8 compared with the PHQ-9 in a large population-based military study which uses self-reported survey data to identify participants who may have depression symptoms. To our knowledge, this is the first study to assess operational differences of the PHQ-8 compared with the PHQ-9 in a large, homogeneous, population-based sample of military men and women. We found the PHQ-8 to be very comparable to the PHQ-9, with only 368 of the 143,705 (0.26%) study participants screening discordantly when using the PHQ-8 compared with the PHQ-9. The reason these 368 participants screened negative using the PHQ-8 and positive using the PHQ-9 was because they only endorsed a total of five depression items on the PHQ-9, one of which was "thoughts that you would be

Table 2
Adjusted odds of screening positive with PHQ-9 and negative with PHQ-8 at baseline, compared with concordant negative and concordant positive PHQ-9 and PHQ-8 groups at baseline.

Baseline characteristics	Screening positive with PHQ-9 and negative with PHQ-8 at baseline							
	Compared with concordant *negative* PHQ-9 and PHQ-8 at baseline				Compared with concordant *positive* PHQ-9 and PHQ-8 at baseline			
	OR	(95% CI)	df	*P*-value	OR	(95% CI)	df	*P*-value
Sex			1	0.9560			1	0.0006
Male	1.00				1.00			
Female	0.99	(0.78, 1.27)			0.65*	(0.51, 0.83)		
Birth cohort			2	0.0230			2	0.4561
Pre-1970	1.00				1.00			
1970–1979	1.37	(0.98, 1.92)			1.16	(0.82, 1.64)		
1980 and later	1.63*	(1.15, 2.31)			1.26	(0.88, 1.79)		
Education			1	0.1417			1	0.5567
Some college or less	1.00				1.00			
Bachelor's or higher	0.66	(0.38, 1.15)			0.85	(0.49, 1.47)		
Marital status			1	0.1258			1	0.0947
Currently married	1.00				1.00			
Not currently married	1.21	(0.95, 1.54)			1.23	(0.97, 1.57)		
Race/ethnicity			2	0.7146			2	0.8355
White non-Hispanic	1.00				1.00			
Black non-Hispanic	1.12	(0.81, 1.55)			1.00	(0.72, 1.38)		
Other	0.96	(0.72, 1.28)			0.92	(0.68, 1.23)		
Military pay grade			1	0.8635			1	0.0735
Enlisted	1.00				1.00			
Officer	1.05	(0.58, 1.90)			1.72	(0.95, 3.13)		
Service branch			3	0.0003			3	0.1606
Army	1.00				1.00			
Navy/coast guard	1.16	(0.88, 1.54)			1.31	(0.98, 1.75)		
Air force	0.51*	(0.36, 0.73)			0.88	(0.61, 1.25)		
Marines	1.07	(0.77, 1.49)			1.15	(0.82, 1.61)		
Component			1	0.3171			1	0.0024
Active duty	1.00				1.00			
Guard/reserve	1.13	(0.89, 1.44)			1.47*	(1.15, 1.89)		
Occupation			2	0.2951			2	0.0505
Combat specialist	1.00				1.00			
Health care	1.08	(0.72, 1.61)			1.12	(0.74, 1.70)		
Other	0.86	(0.66, 1.13)			0.78	(0.60, 1.03)		
Smoking status			2	0.0270			2	0.0115
Never smoker	1.00				1.00			
Past smoker	0.78	(0.59, 1.04)			0.65*	(0.48, 0.87)		
Current smoker	1.18	(0.92, 1.51)			0.80	(0.62, 1.03)		
CAGE/alcohol[a]			1	< 0.0001			1	0.0041
No	1.00				1.00			
Yes	1.74*	(1.39, 2.17)			1.39*	(1.11, 1.74)		
Baseline PTSD[b]			1	< 0.0001			1	< 0.0001
No	1.00				1.00			
Yes	21.55*	(17.24, 26.93)			0.58*	(0.47, 0.72)		
Deployment experience[c]			2	0.7217			2	0.1726
Nondeployed	1.00				1.00			
Deployed without combat	1.03	(0.68, 1.56)			1.33	(0.87, 2.05)		
Deployed with combat	0.90	(0.68, 1.19)			0.85	(0.64, 1.13)		

OR, odds ratio; CI, confidence interval; df, degrees of freedom; DSM-IV-TR, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision; PCL-C, PTSD Checklist-Civilian Version; PTSD, posttraumatic stress disorder.

* Wald chi-square test was significant at the $\alpha = 0.05$ level.

[a] At baseline, participant self-reported ever feeling at least one of the following: (1) need to cut back on drinking, (2) annoyed at anyone who suggested to cut back on drinking, (3) guilty about drinking, and (4) a need for an "eye-opener," or early morning drink.

[b] Participants who screened positive for PTSD symptoms based on PCL-C and DSM-IV-TR criteria at baseline.

[c] Participants who had been deployed and reported witnessing death, trauma, injuries, prisoners of war, or refugees at baseline, were considered to have combat-associated experiences.

better off dead or of hurting yourself in some way" (available on the PHQ-9, but not the PHQ-8). Restriction to the PHQ-8 items only resulted in the endorsement of a total of four depression items, which no longer met criteria for a positive depression screen. The PHQ-8 had almost perfect agreement with the PHQ-9, and it performed similarly when assessing a 5-point or greater increase or decrease in score from baseline to first follow-up. Finally, multivariable analyses identified characteristics associated with being PHQ-9 positive and PHQ-8 negative compared with the two concordant groups, and highlighted the complex relations between PTSD, depression, alcohol abuse, and suicidal thoughts.

Using the PHQ-9 as the gold standard, we observed very high sensitivity and specificity for the PHQ-8 in detecting depression. By design, specificity was 100%, since it is not possible to screen positive for depression on the PHQ-8 and negative on the PHQ-9. Sensitivity was very high for the PHQ-8 at all assessments, with at least 93.7% who screened positive for depression with the PHQ-9 also screening positive using the PHQ-8. Comparisons of the differences in ability between the PHQ-8 and PHQ-9 to detect a 5-point or greater difference in depression score also had very high sensitivity and specificity. These findings are encouraging since this is the first study to our knowledge to examine the

ability of the PHQ-8 to detect a change of magnitude considered to be clinically significant (Kroenke and Spitzer, 2002; Kroenke et al., 2010). The kappa statistics for comparing the ability of these screening tools to detect a 5-point difference in depression scores were extremely high, with each kappa statistic in the "greater than substantial agreement" range of 0.8 to 1.0 (Landis and Koch, 1977). This also provides support for using the PHQ-8 prospectively to detect changes over time in this large, homogeneous, population-based, sample of current and former military personnel.

While there was almost perfect agreement between the PHQ-9 and the PHQ-8, indicating that nearly all participants screened the same using the PHQ-8 and the PHQ-9, there were a small number of participants who were classified as discordant (positive using the PHQ-9 and negative using the PHQ-8). Although this discordant group consists of a very small percentage of the population, there were some statistically significant differences between the discordant and concordant groups. This indicates that those who score positive on the PHQ-9 but negative using the PHQ-8 are different from both the concordant groups in some potentially important ways.

When comparing those who screened PHQ-9 positive/PHQ-8 negative with those screening PHQ-9 negative/PHQ-8 negative or PHQ-9 positive/PHQ-8 positive, different sets of predictors emerged, with only one common factor between the two analyses. In both models, CAGE/alcohol was positively associated with screening PHQ-9 positive/PHQ-8 negative compared with either reference group. This finding likely represents comorbidity between depression, alcohol abuse, and suicidal behavior as previously reported by others (Cornelius et al., 1995; Ganz and Sher, 2009; Henriksson et al., 1993), and it may indicate an advantage associated with screening for suicidal behavior among those with depression, alcohol abuse, especially in the presence of comorbid depression and alcohol abuse.

In this study, quite disparate findings were observed by screening positive for baseline PTSD between the two models shown in Table 2. In the model using concordant negative as the comparison group, those with baseline PTSD were more than 20 times more likely to screen PHQ-9 positive/PHQ-8 negative, while in the model using concordant positive as the comparison group, those with baseline PTSD were significantly less likely to screen PHQ-9 positive/PHQ-8 negative. PTSD has been shown to be associated with suicidal ideation in U.S. military personnel (Guerra and Calhoun, 2011; Jakupcak et al., 2011). This finding reiterates the fact that the PHQ-9 is the preferred tool to use in clinical settings to detect depression, where there is an immediate opportunity to intervene in the event that suicidal thoughts are expressed. One study found that the association between PTSD and suicidal ideation could be fully explained by the presence of depression symptoms (Bryan and Corso, 2011). Also interesting was that another large, longitudinal study of veterans found that comorbid PTSD and depression were associated with lower suicide rates compared with those having only depression (Zivin et al., 2007). Zivin et al. hypothesized that those with comorbid PTSD and depression may have differentially sought more mental health care than those with depression alone, and that being engaged in the health care system may have decreased the risk for suicidal behavior. Undoubtedly, the risk for suicidal behavior in comorbid PTSD and depression deserves further study.

There were some possible limitations to our study. Because of nonresponse to survey invitation, the Millennium Cohort Study may not fully represent the U.S. military; though previous evaluations have shown the cohort to be well representative of military personnel (Ryan et al., 2007; Smith, 2009, 2007a; Wells et al., 2008;). Depression and PTSD were identified based on self-reported symptoms, not clinical diagnoses. However, many U.S. military members do not seek mental health care (Hoge et al., 2004), so the use of electronic medical data and associated diagnoses would likely have led to significant underreporting. Participants self-reported *ever* experiencing combat at baseline, and therefore some experiences reported may not have occurred during deployments in support of the recent operations. Other health and behavioral characteristics were also self-reported and thus subject to potential recall bias; though previous research on this cohort has shown participant data to be reliable (Leardmann et al., 2007; Smith et al., 2007b, 2007c, 2007d, 2007e). Finally, all study participants completed the PHQ-9, and PHQ-8 scores were based upon scoring items A through H of the PHQ-9. It is possible that some participants may have scored differently if they had completed the PHQ-8 rather than extrapolating the PHQ-8 score from the PHQ-9.

There were also several strengths to this study. This was the first study to our knowledge to compare the PHQ-9 and PHQ-8 instruments in a large population-based military cohort. The longitudinal design allowed for prospective follow-up of symptom reporting, and thus the capability to compare the ability to detect a clinically meaningful change by both instruments. The many military, demographic, and health characteristics available for multivariable analyses were a major strength. Additionally, the study population consisted of a large military sample, including members from all service branches, from the Reserves and National Guard, as well as those separated from the military, increasing the generalizability of the findings. Finally, stigma attached with mental disorders such as depression may prevent individuals from seeking care within the Military Health System; therefore, use of confidential questionnaires eliciting self-reported symptoms may have allowed for greater capture of mental health morbidity.

In summary, the PHQ-8 instrument detected nearly all cases of depression captured by the PHQ-9 instrument. High indices of agreement were seen between both measurement instruments. There was no clear-cut pattern of predictors that would distinguish persons with depression based on the PHQ-9 only that would be missed with the PHQ-8 strategy. The one factor that significantly identified PHQ-9 only in both of these comparisons was reporting one or more items on the CAGE/alcohol questions. For population screening, the PHQ-8 instrument has high sensitivity and would capture nearly all the same cases of depression as the PHQ-9 in similar populations to this study.

Disclaimer

This represents report 12–20, supported by the Department of Defense, under work unit no. 60002. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of the Army, Department of the Air Force, Department of Defense, or the U.S. Government. This research has been conducted in compliance with all applicable federal regulations governing the protection of human subjects in research (Protocol NHRC.2000.0007).

Role of funding source

The Millennium Cohort Study is funded through the Military Operational Medicine Research Program of the U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland. Resources from the VA Puget Sound Health Care System supported Dr. Boyko's involvement in this research. The funding organizations had no role in the design and conduct of the study; collection, analysis, or preparation of data; or preparation, review or approval of the manuscript.

Conflict of interest

None reported.

Acknowledgements

In addition to the authors, the Millennium Cohort Study Team includes Melissa Bagnell, MPH; Nancy Crum-Cianflone, MD, MPH; James Davies; Nisara Granado, MPH, PhD; Dennis Hernando; Kelly Jones, MPH; Lauren Kipp, MPH; Michelle Linfesty; Gordon Lynch; Hope McMaster, MA, PhD; Amanda Pietrucha, MPH; Teresa Powell, MS; Amber Seelig, MPH; Besa Smith, MPH, PhD; Katherine Snell; Steven Speigle; Kari Sausedo, MA; Beverly Sheppard; Martin White, MPH; James Whitmer; and Charlene Wong, MPH; from the Deployment Health Research Department, Naval Health Research Center, San Diego, California; Paul Amoroso, from MultiCare Health System Research Institute, Tacoma, Washington; Gary Gackstetter from Analytic Services, Inc., Arlington, Virginia; Tomoko Hooper from the Uniformed Services University of the Health Sciences, Bethesda, Maryland; Margaret A.K. Ryan, MD, MPH, from Naval Hospital, Camp Pendleton, California; and Tyler C. Smith, MS, PhD, from National University, San Diego, California. We thank Scott L. Seggerman from the Management Information Division, Defense Manpower Data Center, Monterey, California, and Michelle LeWark, also from the Naval Health Research Center. We also thank the professionals from the U.S. Army Medical Research and Materiel Command, especially those from the Military Operational Medicine Research Program, Fort Detrick, Maryland. We appreciate the support of the Henry M. Jackson Foundation for the Advancement of Military Medicine, Rockville, Maryland. We are indebted to the Millennium Cohort Study participants, without whom these analyses would not be possible.

References

Bryan, C.J., Corso, K.A., 2011. Depression, PTSD, and suicidal ideation among active duty veterans in an integrated primary care clinic. Psychological Services 8 (2), 94–103.

Corson, K., Gerrity, M.S., Dobscha, S.K., 2004. Screening for depression and suicidality in a VA primary care setting: 2 items are better than 1 item. American Journal of Managed Care 10 (11), 839–845.

Cohen, J.A., 1960. A coefficent of agreement for nominal scales. Educational and Psychological Measurement, 2037–2046.

Cornelius, J.R., Salloum, I.M., Mezzich, J., Cornelius, M.D., Fabrega Jr., H., Ehler, J.G., Ulrich, R.F., Thase, M.E., Mann, J.J., 1995. Disproportionate suicidality in patients with comorbid major depression and alcoholism. The American Journal of Psychiatry 152 (3), 358–364.

Diagnostic and Statistical Manual of Mental Disorders, 2000. Text Revision, fourth ed American Psychiatric Association, Washington, DC.

Dhalla, S., Kopec, J.A., 2007. The CAGE questionnaire for alcohol misuse: a review of reliability and validity studies. Clinical & Investigative Medicine 30 (1), 33–41.

Fann, J.R., Bombardier, C.H., Dikmen, S., Esselman, P., Warms, C.A., Pelzer, E., Rau, H., Temkin, N., 2005. Validity of the patient health Questionnaire-9 in assessing depression following traumatic brain injury. The Journal of Head Trauma Rehabilitation 20 (6), 501–511.

Gray, G.C., Chesbrough, K.B., Ryan, M., Amoroso, P., Boyko, E.J., Gackstetter, G.C., Hooper, T.I., Riddle, J.R., Group. a.t.M.C.S., 2002. The millennium cohort study: a 21-year prospective cohort study of 140,000 military personnel. Military Medicine 167 (6), 483–488.

Glantz, S., Slinker, B., 1990. Primer of Applied Regression and Analysis of Variance. McGraw-Hill, New York, NY.

Ganz, D., Sher, L., 2009. Suicidal behavior in adolescents with comorbid depression and alcohol abuse. Minerva Pediatrica 61 (3), 333–347.

Guerra, V.S., Calhoun, P.S., 2011. Examining the relation between posttraumatic stress disorder and suicidal ideation in an OEF/OIF veteran sample. Journal of Anxiety Disorders 25 (1), 12–18.

Henriksson, M.M., Aro, H.M., Marttunen, M.J., Heikkinen, M.E., Isometsa, E.T., Kuoppasalmi, K.I., Lonnqvist, J.K., 1993. Mental disorders and comorbidity in suicide. The American Journal of Psychiatry 150 (6), 935–940.

Hoge, C.W., Castro, C.A., Messer, S.C., McGurk, D., Cotting, D.I., Koffman, R.L., 2004. Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. The New England Journal of Medicine 351 (1), 13–22.

Jakupcak, M., Hoerster, K.D., Varra, A., Vannoy, S., Felker, B., Hunt, S., 2011. Hopelessness and suicidal ideation in Iraq and Afghanistan War Veterans reporting subthreshold and threshold posttraumatic stress disorder. The Journal of Nervous and Mental Disease 199 (4), 272–275.

Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T., Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. The Journal of Affective Disorders 114 (1-3), 163–173.

Kuehn, B.M., 2009. Soldier suicide rates continue to rise military, scientists work to stem the tide. Jama-Journal of the American Medical Association 301 (11) 1111-+.

Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnositc and severity measure. Psychiatric Annals 32 (9), 1–7.

Kroenke, K., Spitzer, R.L., Williams, J.B., Lowe, B., 2010. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. General Hospital Psychiatry 32 (4), 345–359.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. The Journal of General Internal Medicine 16 (9), 606–613.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

Leardmann, C.A., Smith, B., Smith, T.C., Wells, T.S., Ryan, M.A., 2007. Smallpox vaccination: comparison of self-reported and electronic vaccine records in the Millennium Cohort Study. Human. Vaccine 3, 6.

Oquendo, M., Brent, D.A., Birmaher, B., Greenhill, L., Kolko, D., Stanley, B., Zelazny, J., Burke, A.K., Firinciogullari, S., Ellis, S.P., Mann, J.J., 2005. Posttraumatic stress disorder comorbid with major depression: Factors mediating the association with suicidal behavior. American Journal of Psychiatry 162 (3), 560–566.

Ryan, M.A., Smith, T.C., Smith, B., Amoroso, P., Boyko, E.J., Gray, G.C., Gackstetter, G.D., Riddle, J.R., Wells, T.S., Gumbs, G., Corbeil, T.E., Hooper, T.I., 2007. Millennium Cohort: enrollment begins a 21-year contribution to understanding the impact of military service. Journal of Clinical Epidemiology 60 (2), 181–191.

Spitzer, R.L., Kroenke, K., Williams, J.B., 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. Journal of the American Medical Association 282 (18), 1737–1744.

Spitzer, R.L., Williams, J.B., Kroenke, K., Hornyak, R., McMurray, J., 2000. Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. American Journal of Obstetrics and Gynecology 183 (3), 759–769.

Spitzer, R.L., Williams, J.B., Kroenke, K., Linzer, M., deGruy 3rd, F.V., Hahn, S.R., Brody, D., Johnson, J.G., 1994. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. Journal of the American Medical Association 272 (22), 1749–1756.

Smith, T.C., 2009. The US Department of Defense Millennium Cohort Study: career span and beyond longitudinal follow-up. Journal of Occupational and Environmental Medicine 51 (10), 1193–1201.

Smith, B., Smith, T.C., Gray, G.C., Ryan, M.A., 2007a. When epidemiology meets the Internet: Web-based surveys in the Millennium Cohort Study. American Journal of Epidemiology 166 (11), 1345–1354.

Smith, B., Leard, C.A., Smith, T.C., Reed, R.J., Ryan, M.A., 2007b. Anthrax vaccination in the Millennium Cohort: validation and measures of health. The American Journal of Preventive Medicine 32 (4), 347–353.

Smith, B., Wingard, D.L., Ryan, M.A., Macera, C.A., Patterson, T.L., Slymen, D.J., 2007c. U.S. military deployment during 2001–2006: comparison of subjective and objective data sources in a large prospective health study. Annals of Epidemiology 17 (12), 976–982.

Smith, T.C., Jacobson, I.G., Smith, B., Hooper, T.I., Ryan, M.A., Team, F.T., 2007d. The occupational role of women in military service: validation of occupation and prevalence of exposures in the Millennium Cohort Study. International Journal of Environmental Research and Public Health Research 17 (4), 271–284.

Smith, T.C., Smith, B., Jacobson, I.G., Corbeil, T.E., Ryan, M.A., 2007e. Reliability of standard health assessment instruments in a large, population-based cohort study. Annals of Epidemiology 17 (7), 525–532.

Wells, T.S., LeardMann, C.A., Fortuna, S.O., Smith, B., Smith, T.C., Ryan, M.A., Boyko, E.J., Blazer, D., 2010. A prospective study of depression following combat deployment in support of the wars in Iraq and Afghanistan. The American Journal of Public Health 100 (1), 90–99.

Weathers F.W., Litz B.T., Herman D.S., Huska J.A., Keane T.M. The PTSD Checklist (PCL): reliability, validity, and diagnostic utility. In: Paper presented at: Annual Meeting of International Society for Traumatic Stress Studies, San Antonio, TX; 1993.

Wells, T.S., Jacobson, I.G., Smith, T.C., Spooner, C.N., Smith, B., Reed, R.J., Amoroso, P.J., Ryan, M.A., 2008. Prior health care utilization as a potential determinant of enrollment in a 21-year prospective study, the Millennium Cohort Study. The European Journal of Epidemiology.

Zivin, K., Kim, H.M., McCarthy, J.F., Austin, K.L., Hoggatt, K.J., Walters, H., Valenstein, M., 2007. Suicide mortality among individuals receiving treatment for depression in the Veterans Affairs health system: associations with patient and treatment setting characteristics. American Journal of Public Health 97 (12), 2193–2198.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (DD MM YY)
14 03 12

2. REPORT TYPE
Journal Article

3. DATES COVERED (from – to)
2001-2008

4. TITLE
A Comparison of the PRIME-MD PHQ-9 and PHQ-8 in a Large Military Prospective Study, the Millennium Cohort Study

6. AUTHORS
Wells, Timothy; Horton, Jaime; LeardMann, Cynthia; Jacobson, Isabel; Boyko, Edward

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Commanding Officer
Naval Health Research Center
140 Sylvester Rd
San Diego, CA 92106-3521

5a. Contract Number:
5b. Grant Number:
5c. Program Element Number:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number: 60002

8. PERFORMING ORGANIZATION REPORT NUMBER
12-20

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Commanding Officer
Naval Medical Research Center
503 Robert Grant Ave
Silver Spring, MD 20910-7500

Chief, Bureau of Medicine and Surgery
7700 Arlington Blvd
Falls Church, VA 22042

10. SPONSOR/MONITOR'S ACRONYM(S)
NMRC/BUMED

11. SPONSOR/MONITOR'S REPORT NUMBER(s)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Journal of Affective Disorders (2013), 148, 77–83

14. ABSTRACT

The 9-item scale of the Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PHQ-9) is a validated tool for depression screening. Increasingly, the abbreviated version (PHQ-8) is being used in survey research, which excludes the question on suicidal thoughts due to concerns about inability of researchers to monitor such responses in real time. The performance of the PHQ-8 compared with the PHQ-9 has not been assessed in a population-based sample of military or nonmilitary subjects. The Millennium Cohort comprises 143,705 current and former U.S. military personnel. We conducted a cross-sectional comparison of the PHQ-9 and PHQ-8 and prospective analyses to detect a 5-unit change in these measures. A self-administered questionnaire, conducted at 3-year intervals, included the PHQ-9. Greater than substantial agreement was found between the PHQ-8 and 9 instruments (kappas, 0.966 to 0.974 depending on survey cycle). There was similarly high agreement between the PHQ-8 and 9 in detecting a 5-point increase ($\kappa = 0.987$) or decrease ($\kappa = 0.984$) in score. As excellent agreement was detected between the PHQ-8 and PHQ-9 instruments. The PHQ-8 would capture nearly all the same cases of depression as the PHQ-9 in populations similar to the one in this study.

15. SUBJECT TERMS
depression, longitudinal studies, psychometrics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	18a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Commanding Officer
UNCL	UNCL	UNCL	UNCL	9	18b. TELEPHONE NUMBER (INCLUDING AREA CODE) COMM/DSN: (619) 553-8429